Notes on the Bioinformatics of Gene Patents.

Thomas B. Kepler
Department of Microbiology
Boston University School of Medicine

We previously reported [1] a bioinformatic analysis of claim 5 of US patent 5,747,282 (the '282 patent) showing that this claim, if valid, would be infringed by anyone making or using a large number of 15-mers from genes other than *BRCA1* unless limitations in the scope of the claim were read into the claim by the courts. One could imagine, for example, implicitly limiting the claim to DNA actually derived from the *BRCA1* gene, not read as broadly as the claim language in plain English, which would correspond to any molecule meeting the claim's description. There is no case law one way or the other that would clarify how to interpret the claim language, and until and unless there is, the legal standard is to use the broadest interpretation.

We have extended that work now in three ways that are relevant to the ongoing controversy. We have carried out a thorough investigation of the '282 patent, claims 5 and 6, with respect to prior art, finding that prior art that should have been sufficient to invalidate them certainly existed. The patent prosecution record shows no evidence of the requisite searches that would have shown this being conducted, despite similar searches in 1992 (by a different examiner) having been used to reject claims on the famous NIH "EST Patents" under principal inventor J. Craig Venter. We further show that the claims in '282 are not unique in their structure or

in the actual breadth with which they cover DNA sequences from irrelevant genes. That is, the problem of claims likely to be invalid is not confined to just *BRCA1*, but is shared with other patents that have been licensed for diagnostic use, many of which have been enforced, and yet contain claims of the same type, on short DNA molecules derived from genes. Finally, we demonstrate that DNA oligomers of any length worth protecting for diagnostic purposes stand a surprisingly high chance of being found two or more times in the human genome if they are found at all. Thus, in order to make valid claims on DNA sequences of any practical length, one has to carry out an empirical search for their presence in the relevant DNA databases. Our findings suggest any arbitrary oligonucleotide length short enough to be relevant for hybridization or DNA sequence analysis will include at least some molecules that will encounter problems of demonstrating novelty. Unless those "promiscuous" molecules are culled out, the claims are thus likely to be deemed invalid if challenged. This suggests a heuristic rule relevant for patent examiners and prosecutors: "if you claim a DNA sequence, you have to do the searches to prove it is new."

Results

Short polynucleotide sequences of the type claimed in gene patents and patents that cover genetic diagnostic tests are likely to be found multiple times in the genome.

We extended the previous analysis of US patent 5,747,282 ('282) specifically to address the question of prior art, taking into account the one-year "grace period" in

US law during which the applicants' public disclosure does not defeat patentability. The priority date for the '282 patent and several other Myriad patents is 12 August 1994. Allowing one year prior to this date as the "critical date" for prior art, we collected a sequence library from GenBank by querying nucleotide sequences deposited before 12 August 1993. The library we collected, after removing very short sequences deposited as part of protein-DNA structural studies, contains 108,927 sequences comprising 142,247,638 nucleotides.

Table 1. Original sources of DNA sequences used for comparison to '282 claim 1 sequences.

Database	# sequences
Genbank patents	1321
Genbank non-patent	70022
Genbank (Total)	71343
GenInfo Backbone Database	109
DNA database of Japan	3538
European Molecular Biology Laboratory (EMBL) Data Library	33937

The *BRCA1* sequence of '282 claim 1 and its complement contain 4,170 15-mers that can be found among the DNA sequences that remain.

GenBank also contains sequences collected from prior patent filings, and therefore clearly in the prior art and available for search at the US Patent and Trademark Office even without searching through the other sequences in GenBank. DNA sequences available in GenBank prior to 12 August 1993 included 1321 sequences annotated in the form "Sequence *n* from Patent US *number*" comprising 751,834 total nucleotides. The *BRCA1* sequence of '282 claim 1 and its complement contain 12 15-mers that we found among these sequences. That is, 12 molecules claimed

appear to have already been part of patents with data submitted to GenBank via USPTO more than a year before the patent application was filed, and therefore could have been found by a search of extant sequences associated with patents alone.

The claims on 15-mers from the '282 patent are even broader than indicated by this analysis. We limited our analysis to all 15-mers whose sequence was exactly identical to the *BRCA1* sequence specified in '282. What is claimed, however, is any 15-mer from a DNA sequence that encodes a BRCA1 polypeptide. Because of codon degeneracy, the number of sequences falling under this claim is many times larger than the numbers cited above.

Table 2. Results of search for 15-mers claimed in '282 claim 1.

Source	# exact matches
full collection	2506
reverse complement, full collection	1664
patent sequences	4
patent sequences, reverse complement	8

The scope of the claim thus appears to embrace many molecules already in the prior art a year before patent application.

'282 is not anomalous with respect to broad oligomer claims, nor are such claims limited to older patents. Patent claims on short sequences are found in patents that have been issued within the last 5 years as well. We have not reviewed all patents that might make claims to short DNA fragments, nor would there be any simple way to do so, short of reading claims in thousands of patents. From our limited analysis of 120 patents associated with diagnostic patents, we found that 19 of them (15%)

include such claims. These patents have been granted as recently as 2007 and claim oligomers ranging in size from 8 to 54 nucleotides.

US7214483, the latest patent in this set, was granted in March 2007. This patent claims the gene *KCNQ2*, mutations of which are associated with a form of seizures in newborns. The patent is assigned to the University of Utah and has been licensed exclusively to Athena Diagnostics for testing related to epilepsy.

Claim 8 of '483 claims *An isolated fragment of the DNA of SEQ ID NO: 1, wherein said fragment consists of at least 15 consecutive nucleotides of bases 1315–3232 of SEQ ID NO:1.*
Claim 9 covers *An isolated fragment of the DNA of SEQ ID NO: 1, wherein said fragment consists of at least 8 consecutive nucleotides of bases 1315–3232 of SEQ ID NO:1.*

Claim 8 is for any fragment of length 15; claim 9 is for any fragment of length 8. We used Genbank accession gi|22632021|dbj|BD086411.1| "KCNQ2 and KCNQ3-potassium channel genes mutated in benign familial neonatal convulsion (BFNC) and other convulsions", positions 1315 to 3232, and searched for all 15-mers in this sequence and its complement on each chromosome in the human genome by direct comparison, and obtained the results in Table 3.

There were claimed 15-mers at 16,710 locations in the human genome, excluding the minus strand of chromosome 6, where KCNQ2 itself is found. These hits are located uniformly over the plus and minus strands on all chromosomes. The number of hits for octamers (claim 9) on chromosome 1 alone is 7,657,226. That is, there are

millions of claimed DNA molecules whose sequences would be found on just one of the 24 distinct human chromosomes. Based on this result, and the length of chromosome 1, we estimate that the human genome will contain roughly 87 million such hits.

Table 3. "Hits" and deviation from random sequence variation

chromosome	megabases	hits	expected	excess
1	249.3	1600	883.5	0.81
2	243.2	1406	862.0	0.63
3	198.0	1042	701.9	0.48
4	191.2	934	677.6	0.38
5	180.9	958	641.3	0.49
6	171.1	805	606.5	0.33
7	159.1	869	564.1	0.54
8	146.4	826	518.8	0.59
9	141.2	774	500.5	0.55
10	135.5	840	480.4	0.75
11	135.0	833	478.5	0.74
12	133.9	717	474.5	0.51
13	115.2	441	408.2	0.08
14	107.3	567	380.5	0.49
15	102.5	574	363.4	0.58
16	90.4	695	320.3	1.17
17	81.2	751	287.8	1.61
18	78.1	416	276.8	0.50
19	59.1	681	209.6	2.25
20	63.0	2110	223.4	8.44
21	48.1	223	170.6	0.31
22	51.3	391	181.9	1.15
Y	59.4	143	210.5	-0.32

The problem with 15-mers is clear, though not surprising. Even if the sequence of nucleotides in the human genome were random, one would expect about 95% of all 15-mers to be found in the genome. So, the natural question is whether one can simply make claims on longer oligomers, and if so, how long must they be to ensure uniqueness. By conducting direct searches, we find that the departure from

randomness in the genome is large enough to render any such strategy impracticable without actually doing empirical verification. For any length human gene fragment that would be worth protecting up to at least 400 base pairs, the chance of finding it somewhere else in the human genome is not small enough to make a satisfactory claim without actually performing the search.

The key to this surprising result is that the DNA sequences that are likely to be claimed come themselves from the human genome, and indeed from protein-coding regions particularly apt to have encoded useful motifs during evolution. The "language" of DNA sequences is restrictive enough that a string of, for example, 30 nucleotides that is known to occur at least once in the genome is far more likely to occur a second time than a randomly chosen string of 30 nucleotides is to appear at all. To show this, we performed the following test. For several lengths L from 11 to over 1000 nucleotides, we generate a random set of 100,000 L-mers from the NCBI human RefSeq Gene database. The RefSeq Gene database is a carefully curated set of genes from the genome. By restricting attention to these sequences, we provide the closest match to the conditions encountered in diagnostic DNA patenting, in which sequences primarily from genes rather than from intergenic material are of interest. Repeating the analysis with random oligomers from the whole human genome gives an even stronger departure from expectation under randomness.

Figure 1 shows that the odds ratio for finding an oligomer a second time in the human genome given that it appears at least once. This indicator deviates relatively

little from what is expected under a random-nucleotide string model up to length about 18. From that length on, the odds ratio begins to diverge from expectation of randomness, and continues to do so geometrically. The chance of finding a random 100-mer in the genome is less than 10^{-60}, essentially zero. But the chance of a 100-mer chosen randomly from the human RefSeq Gene database being found twice in the human genome is about 2%. While the departures from randomness of natural DNA sequences are well known (see, e.g., [2]), we are unaware of anyone having addressed the issue with regard to patent practice and are surprised by the proportion of very long oligomers that are repeated in the human genome.

Discussion

We find that several patents with claims on oligomers contain one or more such claims on DNA molecules that on empirical analysis will likely reach well beyond what was discovered by the inventors. These claims extend the reach of patent protection to substantial parts of the human genome, and, indeed, all genomes.

Furthermore, there may be no arbitrary DNA length that reliably defines a sequence as unique to a given gene. Claims on some DNA sequences longer than 20 nucleotides, as granted may nonetheless encounter substantial prior art. This implies that either: (1) claims cannot be interpreted in their plain English meaning and will need to be narrowed in interpretation by case law, or (2) the claims as granted and interpreted in their plain English meaning are invalid.

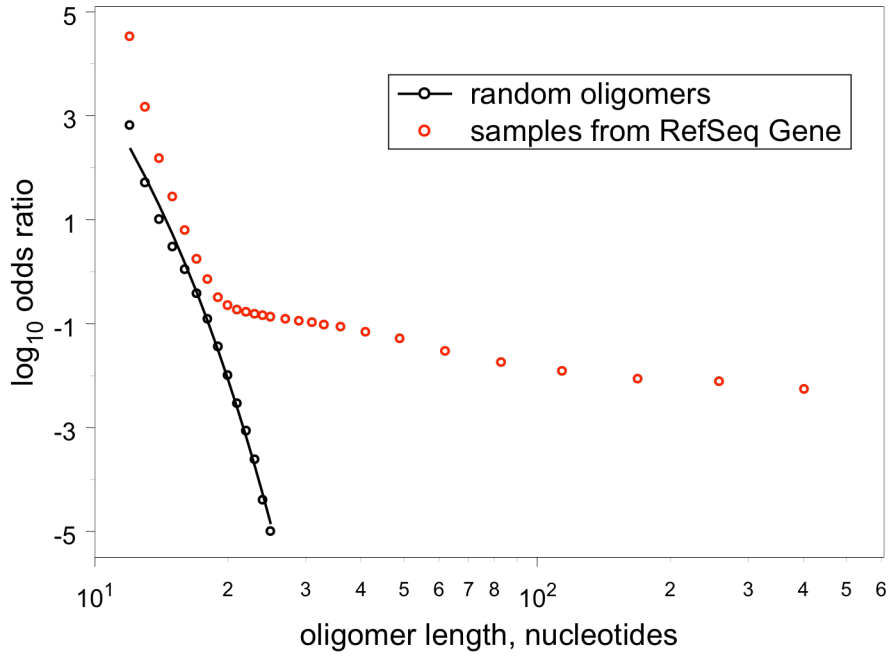

Figure 1. Log of the odds ratio that an oligomer sampled from the human RefSeq Gene database was found twice in the human reference genome given that it appears at least once (red circles) as a function of length. For comparison, we show the log odds ratio that an oligomer produced as a random string of nucleotide symbols chosen at the frequency with they are found in the RefSeq Gene database (black circles). The solid black curve is the linear regression line for the log odds ratio against oligomer

1. Kepler TB, Crossman C, Cook-Deegan R: **Metastasizing patent claims on BRCA1**. *Genomics* 2010, **95**(5):312-314.
2. Csűrös M, Noé L, Kucherov G: **Reconsidering the significance of genomic word frequencies**. *Trends in Genetics* 2007, **23**(11):543-546.